AI-driven security and accountability for autonomous agents

Artificial intelligence is rapidly moving from passive support to active execution. In many organisations, AI is no longer limited to generating text, summarising documents, or supporting analysis. Increasingly, it is embedded in workflows, connected to tools, and authorised to perform tasks with limited human intervention. This shift from “talkers” to “doers” changes the risk landscape materially. It also changes the governance question: the central issue is no longer only what a model can say, but what an agent can access, decide, and do.

As models become more interchangeable, the real control challenge shifts to the harness around them: the permissions, interfaces, context flows, and execution pathways that determine what agents can actually do. This means the governance question is no longer only what a model can say, but what an agent can access, decide, and do. The central issue is therefore not whether agents will be used, but under which conditions they can be deployed safely, responsibly, and with proportionate trust.

This session framed AI not simply as a productivity technology, but as an emerging control challenge. The real question is not whether agents will be used, but under which conditions they can be used safely, responsibly, and with proportionate trust.

Artificial Intelligence’s dual role in cyber security

Erik Beulen
Program Director - Digital Knowledge Institute

AI now plays a dual role in cyber security. On the one hand, it strengthens defence by improving anomaly detection, signal correlation, prediction, response automation, and configuration hardening. On the other hand, it strengthens offence by enabling more persuasive phishing, faster reconnaissance, exploit support, deepfake deception, and better targeting of attacks. The strategic question for organisations is therefore not whether AI has security relevance, but how this dual use should affect risk appetite, governance, and board-level oversight.

Also, autonomous agents are a game changer. These are systems that act towards goals within predefined boundaries by combining perception, planning, action, and adaptation. That functionality creates clear opportunities, but also raises questions about control, ownership, and intervention. If agents can learn from data, change behaviour over time, and execute tasks across systems, responsibility can no longer sit with technology teams alone.

Therefore, emphasise on data ownership, model ownership, and business accountability is important. Data owners control what the system can learn from. Model owners control how the system behaves and adapts. Business, risk, and compliance functions must determine whether a given level

False Clear Ahead - When Agents Derail Your Defenses

Rens van Dongen
Corporate AI Officer – Dutch Railways

Autonomous agents should not be understood as a simple extension of existing AI tools. They represent a structural change in the security model. Traditional software generally operates within relatively stable boundaries, with dependencies resolved before runtime and behaviour constrained by explicit code paths. Unfortunately, agentic systems are different. They act at inference time, ingest changing context, interpret instructions dynamically, and interact with tools, documents, and external systems in ways that are less predictable and harder to verify.

As LLMs become commoditized and merely act as engines, focusing our security efforts on harnessing layers such as Claude Code become essential. In other words, security and trustworthiness depend less on the model brand and more on the surrounding architecture, permissions, isolation, workflow design, and operational controls. This is why not only model behaviour, but also interfaces, tool access, context ingestion, and execution pathways require attention.

There are several core security challenges. The first is the emergence of a probabilistic trusted computing base, as the effective behaviour of the system depends not only on software components, but also on prompts, retrieved content, tool descriptions, environmental signals, and model interpretation. The second is a fuzzy security boundary, because it becomes less clear where the system begins and ends, and therefore where control measures should be applied. The third is dynamic instruction following, with agents continuously reinterpreting goals and context rather than merely executing fixed commands. This creates a larger and more active attack surface than in traditional systems.

Prompt injection is the clearest example of this problem and remains an unresolved frontier issue. In the agentic setting, malicious instructions can be embedded in documents, websites, comments, repositories, or other contextual sources. Because the agent treats context as operational input, these elements can shape reasoning and action directly. The result is that attackers no longer need deep technical sophistication in every case, simple prompting, context manipulation, or content poisoning may be enough to trigger harmful behaviour.

Agentic engineering is important because this is where adoption of AI is advancing fastest. AI is already heavily used in coding, development environments, and adjacent workflows. That makes software engineering the current proving ground for both productivity gains and failure modes. The documents identify four recurring threat categories in this space: 1. promptware, 2. content traps, 3. environment poisoning, and 4. rogue actions. Together, these describe a chain in which manipulated context, compromised environments, unsafe tool use, or excessive autonomy can lead to unintended execution, exfiltration, deletion, or operational disruption.

This problem is not confined to coding. Coding is simply where the tension is most visible today. The broader lesson is that any domain becomes vulnerable when agents combine manipulable inputs, access to sensitive data, and the ability to perform consequential actions. This logic is captured in the “rule of two”, explain that combining too many of these characteristics in one agent creates unacceptable exposure. Security design should therefore separate access, limit privileges, and prevent high-risk combinations wherever possible.

The immaturity of current governance arrangements is another engineering concern, as no enforceable, agent-specific control framework yet exists, and that established frameworks were not designed for autonomous, tool-calling agents. As a result, organisations should prioritise resilience and containment over efficiency optimism. Human oversight remains necessary however this requires some nuance. Effective control does not consist of placing a passive human “in the loop” for every action. It requires meaningful intervention capacity, meaning relevant information, competence, discretion, timely escalation, authority to stop, and organisational accountability have to be in place.

In short, a shift in mindset is much needed. Agentic AI should not be deployed on the assumption that traditional software controls are sufficient. Nor should organisations assume that a capable model automatically implies a trustworthy system. Security depends on design discipline including workload isolation, guarded interfaces, controlled context, monitoring for abuse, strong review points, and clear boundaries on action.

Unfortunately, therefore the overall conclusion is sober but practical, as agents may create value, but without explicit engineering and governance controls, they can just as easily derail the very defences they are meant to strengthen. This is why organisations need to ensure resilient software engineering!

Autonomy & Governance - Agentic AI in the Public Sector

Lüke van den Wittenboer
Data & AI engineer - Atos

AI capability is advancing faster than most organisations can adopt it. Agentic AI not as a distant technology trend, but as a practical organisational challenge already entering all enterprise and public-sector environments. AI can now generate content, reason over documents, use tools, write code, summarise meetings, prepare decisions, and coordinate multi-step tasks. The central issue is therefore no longer whether AI can be useful, but whether work practices, roles, governance, and operating models are evolving quickly enough to capture value safely and consistently.

The widening gap between AI adoption and AI maturity is a concern. Many organisations are experimenting with AI, but shared practices, reusable assets, governance arrangements, and value measurement often lag behind. Knowledge, prompts, agents, context, and lessons learned tend to remain local within teams. At the same time, employees are being asked to “use AI” while their roles, responsibilities, and decision rights are still being defined. This creates friction, as AI activity increases, but organisational learning, accountability, and measurable return on investment remain difficult to prove.

AI value creation is as a change in how work moves through the organisation. AI has the most impact where workflows slow down because of handoffs, coordination effort, documentation, status visibility, information gathering, and process navigation. AI transformation is not only about automation. It is about redesigning workflows, clarifying human roles, and creating an operating model in which humans provide direction, judgment, review, exception handling, stakeholder alignment, and outcome ownership.

The public-sector example of the Flemish Government illustrated this shift in practice. Their citizens faced fragmented access across many websites, service desks, languages, and complex queries, while public administrations face increasing pressure to respond efficiently. The proposed solution is one digital assistant for everyone, acting as a single entry point into multiple specialised agents across domains such as tax, benefits, and care-related services. The intended outcomes are improved citizen experience, reduced contact-centre workload, multilingual 24/7 access to accurate information, lower operational cost, privacy compliance, and scalability.

Such systems only work if governance is built into the design. The Flemish Government approach includes guardrails to block malicious intent, improve answer quality, scope knowledge and responsibility, and ground responses in a knowledge base that is updated daily. It also includes escalation to human hand-over, quality assurance through testing at each deployment or model upgrade and PII masking. This reflects an important lesson: agentic AI in government must be trustworthy by architecture, not merely persuasive in conversation.